Journal Paper

"Preface: Handbook of Research on Big Data Clustering and Machine Learning."

IGI Global book series Advances in Data Mining and Database Management (ADMDM)

(ISSN: 2327-1981; eISSN: 2327-199X)

Fausto Pedro García Márquez
Ingenium Research Group, Universidad de Castilla-La Mancha
FaustoPedro.Garcia@uclm.es

Cite as: Márquez, F. P. G. (2020). Preface. In *Handbook of Research on Big Data Clustering and Machine Learning(pp 17-21)*. IGI Global.

DOI: 10.4018/978-1-7998-0106-1

Preface

Big Data and Management Science has been designed and done to synthesize the analytic principles with business practice and big data. Specifically, the book provides an interface between the main disciplines of engineering/technology and the organizational, administrative, and planning abilities of management. It is complementary to other sub-disciplines such as economics, finance, marketing, decision and risk analysis, etc.

The Advances in Analytics in Big Data synthesizes the analytic principles with Big Data and provides an interface between the main disciplines of engineering/economics and the organizational, administrative, and planning abilities of management. It is also complementary to other disciplines such as finance, marketing, decision and risk analysis. In this book each chapter discusses different topics in Advances in Business Analytics

This book will aim to provide relevant theoretical frameworks and the latest empirical research findings in the area. It will be written for professionals who want to improve their understanding of the strategic role of trust at different levels of the information and knowledge society, that is, trust at the level of the global economy, of networks and organizations, of teams and work groups, of information systems and, finally, trust at the level of individuals as actors in the networked environments.

This book is intended for engineers, economists and researchers who wish to develop new skills in management, or who employ the management discipline as part of their work. The authors of this volume describe their original work in the area or provide material for case studies successfully applying the management discipline in real life cases where is employed Bid Data.

Big data concept became a chief strength of innovation across academics, governments and corporates. Big data comprises massive sensor data, raw and semi-structured log data of IT industries and the exploded quantity of data from social media. Big data need big storage and this volume makes operations such as analytical operations, process operations, retrieval operations, very difficult and time consuming. One way to overcome these difficult problems is to have big data clustered in a compact format. Thus, Chapter 1 discusses the background of big data and clustering. It also discusses the various application of big data in detail. The various related work, research challenges of big data and the future direction would be addressed in this chapter.

Uncertainty is expressed as a situation in which many different outcomes of an option can take place in the decision-making process, but the probabilities of these different outcomes are unknown. When uncertainty is regarded as a surprise and an event in the minds, it can be said that individuals can change the future view. Market, financial, operational, social, environmental, institutional and humanitarian risks and uncertainties are the inherent realities of the modern world. Life is suffused with randomness and volatility; everything momentous that occurs in the illustrious sweep of history, or in our individual lives,

is an outcome of uncertainty. An important implication of such uncertainty is the financial instability engendered to the victims of different sorts of perils. Chapter 2 is intended to explore big data analytics as a comprehensive technique for processing large amounts of data to uncover insights. Several techniques before big data analytics like financial econometrics and optimization models have been used. Therefore, initially these techniques are mentioned. Then, how big data analytics has altered the methods of analysis is mentioned. Lastly, cases promoting big data analytics are mentioned.

Big Data is a critical concept that integrates all kinds of data and plays an important role for strategic intelligence for any modern company. The importance of big data does not revolve around how much data you have, but what you do with it. Big data is now the key for competition and growth for new startups, medium and big enterprises. Scientific research is now on boom using big data. For the astronomers, Sloan Digital Sky Survey has become a central resource. Big data has the potential to revolutionize research and education as well. The aim of Chapter 3 is to discuss the technologies which are pertinent and essential for the big data.

Chapter 4 presents an outline of clustering and Bayesian schemes used in data mining, machine learning communities. Standardize data into sensible groups is preeminent modes of understanding as well as learning. A cluster constitutes set regarding entities which are alike and entities from different clusters are not alike. Representing data by fewer clusters inevitably loses certain fine important information but achieves better simplification. Basically, there is no training stage in clustering; mostly it is used when the classes are not well-known in prior. Bayesian network is one of the best classification method which is frequently used. Generally, Bayesian network is a form of graphical probabilistic representation model consist a set of interconnected nodes, where each node represents a variable and inter-link connection represents a causal relationship of those variables. Belief networks are graph symbolize models that successfully model familiarity estate via transmit probabilistic information to a variety of assumption.

Social media becomes very popular in everyday life. Hence, as a result, the database becomes huge. Therefore, many enterprises are shifting their analytical databases towards cloud instead of high-end proprietary machines and moving towards a cheaper solution. Hence, the concept of MapReduce comes into consideration that provides better scalability, fault tolerance, and flexibility in handling unstructured analytical data.

Chapter 5 shows the importance of GSA, its hybridization and applications in solving clustering and classification problems. In clustering, GSA is hybridized with other optimization algorithms to overcome the drawbacks such as curse of dimensionality, trapping in local optima and limited search space of conventional data clustering algorithms. GSA is also applied to classification problems for pattern recognition, feature extraction, and increasing classification accuracy.

With the importance of forecasting in businesses, a wide variety of methods and tools has been developed over the years to automate the process of forecasting. However, an unintended consequence of this tremendous advancement is that forecasting has become more and more like a black box function. Thus, a primary goal of Chapter 6 is to provide a clear understanding of forecasting in any application contexts with a systematic procedure for practical forecasting through step-by-step examples. Several methods are presented and the authors compare results to what were the typical forecasting methods including regression and time series in different software technologies. Three case studies are presented: simple supply forecasting, homicide forecasting, and demand forecasting for sales from Walmart.

Machine learning, big data, and high dimensional data are the topics we hear about frequently these days and some even call them the wave of the future. Therefore, it is important to use appropriate statistical models, which have been established for many years and their efficiency have already been evaluated, to

contribute into advancing machine learning, which is a relatively newer field of study. Different algorithms that can be used within machine learning, depending on the nature of the variables, are discussed and appropriate statistical techniques for modeling them are presented in Chapter 7.

Regression model is an important tool for modeling and analyzing data. In Chapter 8, the proposed model comprises of three phases. First phase concentrates on sampling techniques to get best sample for building the regression model. Second phase is to predict the residual of Logistic Regression (LR) model using time series analysis method- Autoregressive. Third phase is to develop Enhanced Logistic Regression (ELR) model by combining the both LR model and Residual Prediction (RP) Model. The empirical study is carried out to the study the performance of the ELR model using large diabetic dataset. The results show that ELR model has higher level of accuracy than the traditional Logistic Regression model.

The aim of parametric regression models like linear regression and nonlinear regression are to produce a reasonable relationship between response and independent variables based on the assumption of linearity and predetermined nonlinearity in the regression parameters by finite set of parameters. Nonparametric regression techniques are widely-used statistical techniques and they are not only relax the assumption of linearity in the regression parameters, but they also do not need a predetermined functional form as nonlinearity for the relationship between response and independent variables. It capable of handling higher dimensional problem and sizes of sample than regression that considers parametric models because the data should provide both the model building and the model estimates. For this purpose, in Chapter 9, firstly, PRSS problems for MARS, ADMs and CR will be constructed. Secondly, the solution of the generated problems will be obtained with CQP, one of the famous methods of convex optimization, and these solutions will be called CMARS, CADMs and CKR respectively.

The expeditious increase in the adoption of social media over the last decade, determining and analyzing the attitude and opinion of masses related to a particular entity has gained quite an importance. With the landing of the Web (2.0), many internet products like Blogs, Community Chatrooms, Forums, Microblog are serving as a platform for people to express themselves. Such opinion is found in the form of messages, user-comments, news articles, personal blogs, tweets, surveys, status updates etc. With sentiment analysis, it is possible to eliminate the need to manually going through each and every user comment by focusing on the contextual polarity of the text. Analyzing the sentiments could serve a number of applications like advertisements, recommendations, quality analysis, monetization provided on the Web services, real-time analysis of data, analyzing notions related to candidates during election campaign, etc. This is analysed in Chapter 10.

Chapter 11 examines the performance of liquidity-adjusted risk modeling in obtaining optimum and coherent economic-capital structures, subject to meaningful operational and financial constraints as specified by the portfolio manager. Specifically, the chapter proposes a robust approach to optimum economic-capital allocation, in a Liquidity-Adjusted Value at Risk (L-VaR) framework. This chapter expands previous approaches by explicitly modeling the liquidation of trading portfolios, over the holding period, with the aid of an appropriate scaling of the multiple-assets' L-VaR matrix along with GARCH-M technique to forecast conditional volatility and expected return. Moreover, in this chapter, the authors develop a dynamic nonlinear portfolio selection model and an optimization algorithm, which allocates both economic-capital and trading assets by minimizing L-VaR objective function. The empirical results strongly confirm the importance of enforcing financially and operationally meaningful nonlinear and dynamic constraints, when they are available, on the L-VaR optimization procedure.

A data-driven stochastic program for bi-level network design with hazardous material (hazmat) transportation is proposed in Chapter 12. In order to regulate the risk associated with hazmat transpor-

tation and minimize total travel cost on interested area under stochasticity, a multi-objective stochastic optimization model is presented to determine generalized travel cost for hazmat carriers. Since the bi-level program is generally non-convex, a data-driven bundle method is presented to stabilize solutions of the proposed model and reduce relative gaps between iterations. Numerical comparisons are made with existing risk-averse models. The results indicate that the proposed data-driven stochastic model becomes more resilient than others in minimizing total travel cost and mitigating risk exposure. Moreover, the trade-offs among maximum risk exposure, generalized travel costs and maximum equitable risk spreading over links are empirically investigated in this chapter.

Chapter 13 uses "Mobile Kukan Toukei™" (mobile spatial statistics) to collect the location data of mobile phone users in order to count the number of visitors at specific tourist destinations and examine their characteristics. Mobile Kukan Toukei is statistical population data created by a mobile phone network. It is possible to estimate the population structure of a region by gender, age, and residence using this service of the company. The locations and characteristics of the individuals obtained herein are derived through a non-identification process, aggregation processing, and concealment processing. Therefore, it is impossible to identify specific individuals. This chapter attempts to identify the number of visitors in different periods and their characteristics based on the location data of mobile phone users collected by the mobile phone company. In addition, it also attempts to demonstrate an alternative method to more accurately infer the number of visitors in specific areas.

The unprecedented growth in the amount and variety of data it can be stored about the behaviour of customers has been parallel to the popularization and development of machine learning algorithms. This confluence of factors has created the opportunity of understanding customer's behaviours and preferences in ways that were undreamt of in the past. Chapter 14, the authors study the possibilities of different state-of-the-art machine learning algorithms for retail and smart tourism applications, which are domains that share common characteristics, such as contextual dependence and the kind of data that can be used to understand customers. They explore how supervised, unsupervised and recommender systems can be used to profile, segment and create value for customers.

Industrial robotics is constantly evolving, with installation forecast of about 2 million new robots in 2020. It is necessary to be more efficient in the processes, producing more in less time, which implies reducing the time of penalizing breakdowns of the plants. The predictive maintenance focused on industrial robots is beginning to be applied more, but its possibilities have not yet been fully exploited. Chapter 15 focuses on the applications offered by inertial sensors in the field of industrial robotics, specifically the possibility of measuring the "real" rotation angle of a robotic arm and comparing it with its own system of measure. This could determinate the need to make actions plans to extend the life of industrial robots and avoid unwanted stops of production processes, which could only be solved through corrective actions. The study will focus on the measurement of the backlash existing in the gearbox of the axis of a robot. Data received from the sensor will be analysed using the Wavelet Transform, and the mechanical state of the system could be determined. The introduction of this sensing system is safe, dynamic and non-destructive, and it allows to perform the measurement remotely, in the own installation of the robot and in working conditions. The data of the sensor can be stored to determine pattern of movements and compare them in the future with the current values. All these features allow to use the device in different predictive functions.

Chapter 16 discussed the phenomenon of call masking, and other related infractions have assumed frightening dimension in Nigeria. Apart from depriving the government and telecoms companies of huge revenue, the sharp practices also constitute security threat to the nation. In a bid to curb the men-

ace, the Nigerian Communications Commission, the industry regulator, had to suspend six interconnect exchange licenses in February 2018 and bar 750,000 lines belonging to 13 operators from the national network suspected to have been involved in the criminal act. However, in spite of the measures taken by NCC, the sharp practices have continued unabated. It is against this backdrop that this chapter proffers solutions and recommends ways to nip the infractions in the bud and save the telecoms industry from imminent collapse.

Chapter 17 focuses on Discrete Firefly Optimization Algorithm (FA) based Microarray Data which is a meta-heuristic, bio-inspired, optimization algorithm based on the flashing behaviour of fireflies, or lighting bugs. Its primary advantage is the global communication among the fireflies, and as a result, it seems more effective for triclustering problem. This chapter aims to render a clear description of a new Firefly Algorithm (FA) for optimization of tricluster applications. this research work proposes Discrete Firefly Optimization based Triclustering model first time to find the highly correlated tricluster from microarray data. This model is reliable and robust triclustering model because of efficient global communication among the swarming particles called fireflies.

In Chapter 18, machine learning techniques are applied to examine consumer food choices, specifically purchasing patterns in relation to fresh fruit and vegetables. This product category contributes some of the highest profit margins for supermarkets, making understanding consumer choices in that category important not just for health, but also for economic reasons. Several unsupervised and supervised machine learning techniques, including hierarchical clustering, latent class analysis, linear regression, artificial neural networks, and deep learning neural networks are illustrated using the Nielsen Consumer Panel Dataset, a large and high-quality source of information on consumer purchases in the United States. The main finding from the clustering analysis is that households that buy less fresh produce are those with children – an important insight with significant public health implications. The main outcome from predictive modelling of spending on fresh fruit and vegetables is that contrary to expectations, neural networks failed to outperform linear regression models.

Finally, Chapter 19 studies the urban spatial data as the source of information in analysing risks due to natural disaster, evacuation planning, risk mapping and assessments etc. Global Positioning System (GPS) is a satellite based technology which is used to navigate on earth. It enables tracking of stationary and non-stationary objects in real world. Geographical Information System (GIS) is a software system that facilitates software services to mankind in various application domains such as agriculture, ecology, forestry, geomorphology analysis in earthquake and landslides, laying of underground water pipe connection and demographic studies like population migration, urban settlements etc. Spatial and temporal analysis of such human activities involves aggregation of spatial and temporal factors. Further, these factors act as prime elements in decision making on human activities when they are related. Thus, spatial and temporal relations of real time activities can be analysed to predict the future activities like predicting places of interest. Time analysis of such activities helps in personalisation of activities or development of recommendation systems which could suggest places of interest. Thus, GPS mapping with data analytics using GIS would pave way for commercial and business development in large scale.

Fausto Pedro García Márquez
University of Castilla-La Mancha, Spain